

Data Science (CS 839), Spring 2019

Project Stage 1 – Named Entity Recognition

Team Members:

- Aditya Rungta (aditaker@cs.wisc.edu)
- Arjun Balasubramanian (balarjun@cs.wisc.edu)
- Rohit Kumar Sharma (rsharma@cs.wisc.edu)

Dataset:

A dataset of news articles from BBC was used for the project. About 386 documents from Entertainment section were extracted from the above link.

Named Entity Annotation and Examples:

We chose to extract location for the project. We annotated 330 of the documents and used 300 of them for training and testing. For annotation, we marked the locations with `<loc></loc>` tags. Few examples are shown below:

...Hailey emigrated to `<loc>Canada</loc>` in 1947...
...when he arrived in `<loc>London</loc>` from `<loc>Polynesia</loc>`...
...A `<loc>US</loc>` television network will...

The number of documents chosen for Dev (set I) and Test (set J) set splits each with the total number of annotations is shown in the table below:

| Dataset | I | J | Total |
|-------------------------|-----|-----|-------|
| Total Documents | 200 | 100 | 300 |
| Total location mentions | 912 | 418 | 1330 |

Features Used:

A 58-dimensional feature vector was used to represent the candidate tokens (substrings of up to length two words) for learning. Features chosen for each dimension are:

1. Is the first letter of the token in upper-case?
2. Does the prefix belong to ['at', 'in', 'of', 'the', 'to', 'across', 'from']?
3. The number of words comprising the token.
4. Does suffix start with an upper-case letter?
5. Does prefix start with an upper-case letter?
6. Does prefix belong to ["Mr", "Mrs", "Ms", "Dr", "Prof"].
7. 50-dimensional pretrained GloVe [1] word vectors.

Results:

We performed 4-fold Cross Validation on the dev set (set I) to perform the training. All the below mentioned metrics are the average of 1-fold used for cross validation when trained with the remaining 3 folds.

Before applying any post processing rules, the following metrics were observed:

| Model | Precision | Recall | F1-Score |
|-------------------------|-----------|--------|----------|
| Random Forest | 0.770 | 0.801 | 0.785 |
| Decision Tree | 0.990 | 0.199 | 0.332 |
| Logistic Regression | 0.823 | 0.727 | 0.772 |
| Support Vector Machines | 0.916 | 0.875 | 0.894 |

Support Vector Machines (SVM) model performed the best with an F1-score of 0.894. So, we used SVM for the rest of the tasks.

Before rule-based post processing, SVM reported the following results on the test set (set J):

| | Precision | Recall | F1-Score |
|--------------------------------|-----------|--------|----------|
| Support Vector Machines | 0.913 | 0.876 | 0.894 |

Rule-based Post Processing:

After observing some false negatives, we added a whitelist rule for post processing. The whitelist consists of ['New York', 'Los Angeles', 'San Francisco'].

After applying this rule, the following results were observed on the test set (set J):

| | Precision | Recall | F1-Score |
|--------------------------------|-----------|--------|----------|
| Support Vector Machines | 0.914 | 0.888 | 0.901 |

References:

[1] Jeffrey Pennington and Richard Socher and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. *Empirical Methods in Natural Language Processing (EMNLP)*